Jennifer Yoon memo, date Feb 5, 2025

Folder contents, ***12-lead dataset***s, Google Drive shared.

Jan 23, 2025 Rodrigo uploaded 6 folders.

But **none** of them actually have data files, not even .zip or .tar.

Only papers and some have links, 2 .md text files, paper memos Rodrigo saved.

All are from **year 2020** Physionet CinC Challenge. All data is from there.

Took a long time, 6 hrs. Downloaded all dataset from source paper links.

All saved to local computer, repo data folder, part of .gitignore (not uploaded to Github).

All zip files extracted, examined for errors, organized.

Active datasets copied to “main” folder inside this folder, ***12-lead datasets*** .

All datasets are exact duplicates of datasets found in 2021 CinC Challenge, link on source website. Only missing from this dataset, Shandong Provincial Hospital SPH 25K dataset.

SPH dataset separately downloaded. “records.tar.gz” zip file.

Uncompressed files in “extracted\_files” folder.

—-------------------------------------------

1) A large scale 12-lead ...

I think this is 2nd Chinese dataset Jason mentioned.

Thus, we collected and disseminated this novel database that contains

12-lead ECGs of \*\*\*10,646 patients with a 500 Hz sampling rate\*\*\* that features 11 common rhythms and

67 additional cardiovascular conditions, all labeled by professional experts. The dataset consists of

\*\*\*10-second, 12-dimension ECGs and labels\*\*\* for rhythms and other conditions for each subject. The

dataset can be used to design, compare, and fine-tune new and classical statistical and machine

learning techniques in studies focused on arrhythmia and other cardiovascular conditions.

2018 Shaoxing Medical & Hygiene Research Grant

10,000 patients, new database, Shaoxing People's Hospital, 12-lead, 11 rhythms and 67 conditions (classifications).

Also in folder "A large-scale multi-label 12-lead"

code for reading h5 data format.

PDF of same paper, A large scale ... folder

2) PTB Diagnostic ECG Database - duplicate

markdown file only, nothing new

3) PTB-XL, a large publicly available ... New

To address these issues, we put

forward PTB-XL, the to-date largest freely accessible clinical 12-lead ECG-waveform dataset comprising

\*\*\*21,837 records from 18,885 patients of 10 seconds length\*\*\*. The ECG-waveform data was annotated by

up to two cardiologists as a \*\*\*multi-label dataset\*\*\*, where diagnostic labels were further aggregated into

super and subclasses. the dataset covers a broad range of diagnostic classes including, in particular,

a large fraction of healthy records. the combination with additional metadata on demographics,

additional diagnostic statements, diagnosis likelihoods, manually annotated signal properties as

well as suggested folds for splitting training and test sets turns the dataset into a rich resource for the

development and the evaluation of automatic ECG interpretation algorithms.

4) St. Petersburg - small, not useful

The original records were collected from patients undergoing tests for coronary artery disease (17 men and 15 women, aged 18-80; mean age: 58).

This database consists of 75 annotated recordings extracted from 32 Holter records. Each record is 30 minutes long and contains 12 standard leads, each sampled at 257 Hz,

5) Physionet 2020 Challenge - New.

Does not include Shandong Provincial Hospital 2020 dataset.

All others are exact duplicates of datasets found in 2021 year CinC challenge, link from Shandong paper, source 2021 CinC.

Classification of 12-lead ECGs: the PhysioNet/Computing in

Cardiology Challenge 2020

Also has "papers" folder w 67 linked PDFs, Challenge participants.

6 hospitals from 4 countries. 66,000 recordings, 43,000 were public recordings, rest were hidden (test set). 2020 Physionet Challenge.

"training" has datasets, definitely duplicate St Petersburg, PTB-XL, 2018 Shaoxing??

7.58 GB full folder.

Other folder just has text, may be notes Rodrigo wrote. No data. PDF papers some.

This work addresses these issues by providing a standard, multi

institutional database and a novel scoring metric through a public competition: the PhysioNet/ Computing in Cardiology Challenge 2020.

Approach: A total of \*\*\*66,361 12-lead ECG recordings\*\*\* were sourced from six hospital systems

from \*\*\*four countries across three continents\*\*\*; \*\*43,101 recordings were posted publicly with a focus

on 27 diagnoses\*\*\*. For the first time in a public competition, we required teams to publish open

source code for both training and testing their algorithms, ensuring full scientific reproducibility.

Main results: A total of 217 teams submitted 1395 algorithms during the Challenge,

representing a diversity of approaches for identifying cardiac abnormalities from both academia and industry. As with previous Challenges, high-performing algorithms \*\*\*exhibited significant drops

(⪅ 10%) in performance on the hidden test data\*\*\*.

Significance: Data from diverse institutions allowed us to assess algorithmic generalizability. A

novel evaluation metric considered different misclassification errors for different cardiac

abnormalities, capturing the outcomes and risks of different diagnoses. Requiring both trained

models and code for training models improved the generalizability of submissions, setting a new

bar in reproducibility for public data science competitions.